Convergence diagnostics for MCMC

Vivekananda Roy

Iowa State University

Mainly based on
Roy, V. (2020) Convergence diagnostics for Markov chain Monte Carlo *Annual Review of Statistics and Its Application*, 7 387-412

Measuring the quality of MCMC output workshop
BayesComp-ISBA
October 6, 2021

Monte Carlo methods

We have $u : \mathbb{R}^d \to [0, \infty)$ satisfying $\int_{\mathbb{R}^d} u(y)dy < \infty$.
Define the density function $\pi(x) = u(x)/\int_{\mathbb{R}^d} u(y)dy$.
We Want to know:

$$E_\pi g = \int_{\mathbb{R}^d} g(x)\pi(x)dx = \int_{\mathbb{R}^d} g(x)u(x)dx \Big/ \int_{\mathbb{R}^d} u(y)dy,$$

but both numerator and denominator are analytically intractable.
Ordinary Monte Carlo is the method of using IID simulations $X_1, \ldots, X_n$ from π to approximate expectations by sample averages

$$\overline{g}_n = \frac{1}{n}\sum_{i=1}^{n} g(X_i).$$

By the law of large numbers (LLN), if $\mathrm{E}_\pi|g| < \infty$, $\overline{g}_n \stackrel{as}{\to} E_\pi g$, as $n \to \infty$.
By CLT, if $\mathrm{E}_\pi g^2 < \infty$,

$$\sqrt{n}\Big(\overline{g}_n - \mathrm{E}_\pi g\Big) \stackrel{d}{\to} \mathrm{N}(0, \lambda_g^2).$$

$$s_g^2 = \frac{1}{n}\sum_{i=1}^{n}(g(X_i) - \overline{g}_n)^2.$$

How large should n be?

Asymptotic 95% CI for $\mathrm{E}_\pi g$: $\overline{g}_n \mp 1.96 s_g/\sqrt{n}$.

Markov chain Monte Carlo

We Want to know:

$$E_\pi g = \int_{\mathbb{R}^d} g(x)\pi(x)dx = \int_{\mathbb{R}^d} g(x)u(x)dx \Big/ \int_{\mathbb{R}^d} u(y)dy.$$

Ordinary Monte Carlo is the method of using IID simulations $X_1, \ldots, X_n$ from π to approximate expectations by sample averages

$$\overline{g}_n = \frac{1}{n}\sum_{i=1}^{n} g(X_i).$$

Markov chain Monte Carlo (MCMC) replaces IID simulations with realizations $X_0, X_1, \ldots, X_n$ of a Markov chain that has stationary distribution π and is appropriately irreducible.

By SLLN for Markov chains, if $\mathrm{E}_\pi|g| < \infty$, for any initial X_0, $\overline{g}_n \stackrel{as}{\to} E_\pi g$, as $n \to \infty$, that is, the sample mean *converges* to the population mean.

Also, if f_n denotes the density of X_n, then $\int_{\mathbb{R}^d} |f_n(x) - \pi(x)|dx \downarrow 0$ as $n \to \infty$, that is, the chain converges to the target π.

The further the initial distribution (of X_0) from π, the longer it takes for X_n to approximate π.

Where to start collecting samples and when to stop?

We have $\int_{\mathbb{R}^d} |f_n(x) - \pi(x)| dx \downarrow 0$ and $\overline{g}_n \stackrel{\text{as}}{\to} E_\pi g$ as $n \to \infty$.

Find the smallest n' such that

$$\frac{1}{2} \int_{\mathbb{R}^d} |f_{n'}(x) - \pi(x)| dx < 0.01.$$

Every subsequent draw approximately follows π. That is, the chain has approximately reached stationarity. The above n' is referred to as the honest value for *burn-in* (Jones and Hobert 2001).

Letting $\bar{g}_{n',n} \equiv \sum_{i=n'+1}^{n} g(X_i)/(n-n')$, if

$$\sqrt{n}(\bar{g}_{n'n} - E_\pi g) \stackrel{d}{\to} N(0, \sigma_g^2) \text{ as } n \to \infty,$$

where $\sigma_g^2 \equiv \text{Var}_\pi(g(X_0)) + 2\sum_{i=1}^{\infty} \text{Cov}_\pi(g(X_0), g(X_i)) < \infty$, and if $\widehat{\sigma}_{g,n}$ is a consistent estimator of σ_g, then a 95% CI for $E_\pi g$ is

$$\bar{g}_{n',n} \mp 1.96 \widehat{\sigma}_{g,n}/\sqrt{n-n'}.$$

Fixed and relative width stopping rules

The fixed-width stopping rule (FWSR) terminates the simulation the first time that

$$t_* \frac{\widehat{\sigma}_{g,n}}{\sqrt{n}} + \frac{1}{n} \leq \epsilon.$$

The relative standard deviation FWSR (SDFWSR) terminates the simulation when

$$t_* \frac{\widehat{\sigma}_{g,n}}{\sqrt{n}} + \frac{1}{n} \leq \epsilon \widehat{\lambda}_{g,n},$$

where $\widehat{\lambda}_{g,n}$ is a consistent estimator of λ_g.

The SDFWSR is equivalent to stopping the simulation after reaching certain *effective sample size*, defined as $n\lambda_g^2/\sigma_g^2$.

- Gelman-Rubin diagnostic: $\hat{R}$
- Geweke's statistic
- Heidelberger & Welch's statistic
- Raftery-Lewis diagnostic
- Kernel density based methods
- Graphical tools (trace plots, autocorrelation function plots)

An exponential target distribution

Let $\pi(x) = \exp(-x),\ x > 0$.
Consider independence Metropolis samplers with $\text{Exp}(\theta)$ proposals.
For $\theta < 1$,

$$\frac{1}{2}\int_{\mathcal{X}} |f_n(x) - \pi(x)| dx \le (1-\theta)^n.$$

If $\theta = 0.5$, $n' = \lceil \log(0.01)/\log(0.5) \rceil = 7$.

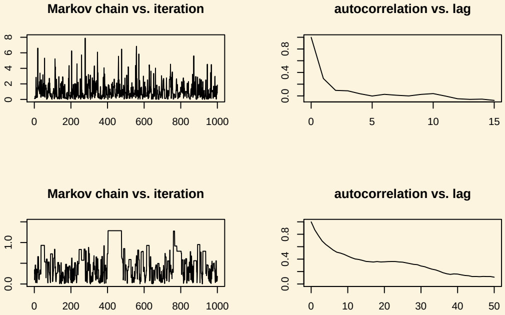

Figure: Trace (left panels) and autocorrelation function (right panels) plots of the independence Metropolis chains (top row, $\theta = 0.5$; bottom row, $\theta = 5$) for the exponential target example.

An exponential target distribution

Let $\pi(x) = \exp(-x),\ x > 0$. Consider estimation of $E_\pi X = 1$.
By FWSR, for $\theta = 0.5$ starting at $X_8 = 0.1545$, it takes $n^* = 323,693$ iterations to achieve the cutoff 0.005. The cut off value for ESS is 153,658 iterations.

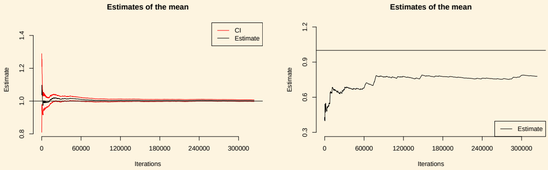

Figure: The left plot shows the running estimates of the mean with confidence interval for $\theta = 0.5$. Running mean plot for $\theta = 5$ is given in the right panel. The horizontal line denotes the truth.

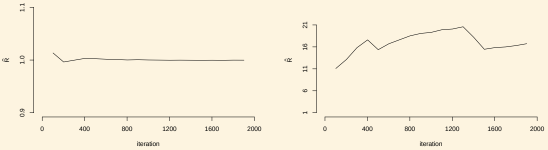

Figure: Iterative $\hat{R}$ plot (from four parallel chains) for the independence chains (left plot $\theta = 0.5$, right plot $\theta = 5$). In the left plot, the PSRF reaches below the cutoff (1.1) before 100 iterations, leading to premature termination of the chain.

A sixmodal target distribution

Let

$$\pi(x, y) \propto \exp\left(\frac{-x^2}{2}\right) \exp\left(\frac{((\csc\ y)^5 - x)^2}{2}\right), \ -10 \leq x, y \leq 10.$$

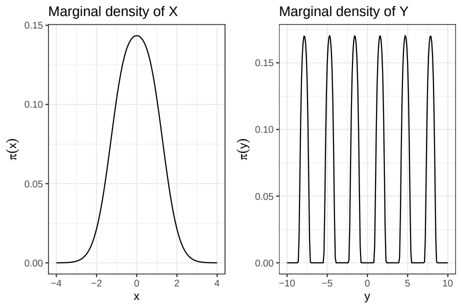

Figure: Marginal densities of X and Y in the sixmodal example.

A sixmodal target distribution

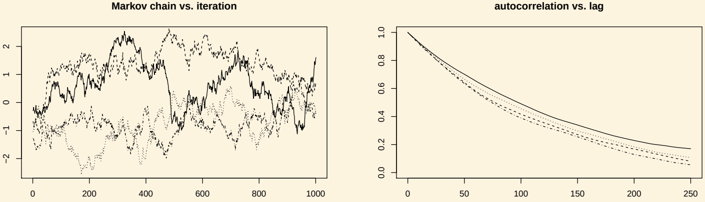

Figure: Trace (left panel) and ACF (right panel) plots of the X marginal of the four chains.

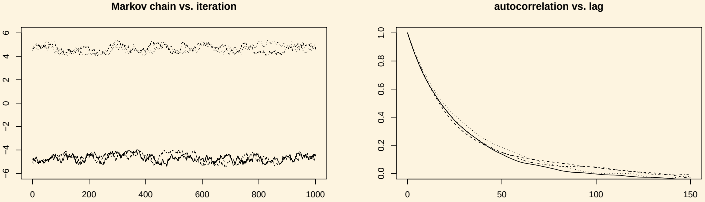

Figure: Trace (left panel) and ACF (right panel) plots of the Y marginal of the four chains. Unlike for the X marginal chains, trace plots of some of the Y marginal chains do not have any overlap demonstrating divergence of the Markov chains.

A sixmodal target distribution

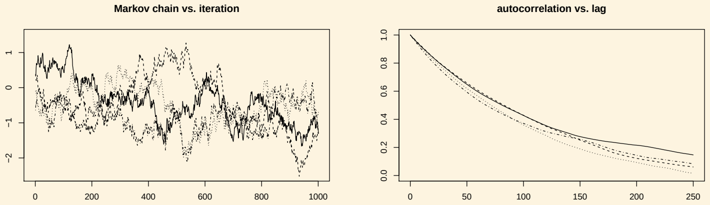

Figure: Trace (left panel) and ACF (right panel) plots of the X marginal of the four chains.

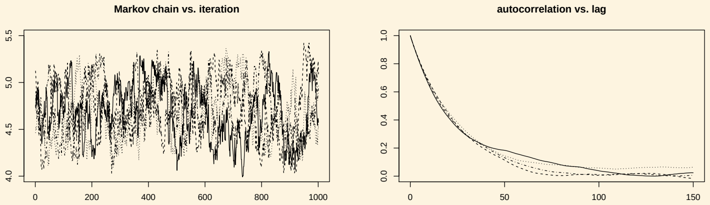

Figure: Trace (left panel) and ACF (right panel) plots of the Y marginal of the four chains. The large amount of overlap between the trace plots of the four marginal chains fails to indicate non-convergence of the Markov chains to stationarity.

Summary

- Where to start collecting samples and when to stop the simulation are related to two convergence concepts.
- Reviewed different popular diagnostics for MCMC convergence.
- Empirical diagnostics may prematurely terminate the simulation.
- In the presence of multiple modes, if the chains are not run long enough, empirical diagnostics may fail to detect the non-convergence.